Attorney Docket No. 07414-0025

UNITED STATES PATENT APPLICATION

OF

Hugh J. PASIKA
Yuandan LOU
David P. HOLDEN

FOR

METHODS, SYSTEMS, AND ARTICLES OF MANUFACTURE FOR EVALUATING BIOLOGICAL DATA

This application claims the benefit of U.S. Provisional Application Serial No. 60/227,556, filed on August 23, 2000. This application incorporates by reference all of the disclosure of U.S. Serial No. 60/227,556.

FIELD OF THE INVENTION

This invention relates generally to data processing systems and, in certain embodiments, allele calling algorithms.

BACKGROUND OF THE INVENTION

There are many techniques for analyzing nucleic acid information. For example, certain techniques involve studying genetic polymorphisms. A polymorphism involves difference in a given portion of a nucleic acid sequence in different individuals within a population. Such polymorphisms may occur in regions in which nucleic acids do not encode proteins. In such regions, often there are large numbers of repeats of a given short sequence. For example, there may be regions of multiple repeats of a given dinucleotide (such as GC or CA), trinucleotides, or larger repeat units. The larger repeat regions (larger number of nucleotide bases within a repeated motif) have often been referred to as "minisatellites." The smaller repeat regions (1, 2, 3, 4, 5, or 6 nucleotides within a repeated motif) have often been referred to as "microsatellites" or "short tandem repeats (STR's)." Through evolution, individuals often vary in the number of repeats at a given locus.

Such repeat regions can serve as genetic markers since individuals can vary in the number of repeats at a given locus (location) or at many loci (locations). Each different form at a given

locus is known as an allele. These differences at a given position can serve as genetic markers that are useful for many purposes including positively identifying an individual from genetic material based on the unique genetic pattern of such an individual. Also, variations between individuals may signify predisposition to a disease or other genetic conditions.

Thus, much effort has been focused on positively identifying particular alleles at given genetic loci. For example, methods of determining the number of dinucleotide repeats at a given locus include use of PCR to amplify the regions in question. One uses primers to locate and initiate amplification of a particular loci in a sample. After the amplification, one determines the particular alleles at a given locus in the sample by determining the fragment length of the amplified material. By determining the fragment length, one can determine the number of dinucleotide repeats at that location. Thus, the particular allele at that locus is identified.

Artifacts, however, can be created in the process, which may render difficult accurate determination of the actual allele at a given locus. These artifacts may be a result of PCR stutter, which can result from mistakes in amplification of the repeated nucleotides in the region being studied. Specifically, the polymerase in the PCR reaction may slip and miss one or more of the repeat units that are present in the studied nucleic acid region. In addition, an extra A nucleotide may be added during amplification. Thus, when PCR stutter and/or plus A distortion occurs, the amplification products typically will include not only the correct amplified allele, but also shorter repeats missing one or more of the repeat units of the allele. In fact, the data may show multiple peaks of various lengths where the data should reflect only one length.

SUMMARY OF THE INVENTION

Methods, systems, and articles of manufacture consistent with certain embodiments of the present invention overcome the shortcomings of conventional allele calling algorithms by providing a committee machine that receives calls from several allele calling algorithms. By receiving calls from multiple allele calling algorithms, each employing a different calling philosophy, the committee machine makes calls containing a high level of confidence over a variety of conditions and transmits statistically meaningful confidence values to the user.

According to certain embodiments of the invention, unique calling algorithms are also provided.

BRIEF DESCRIPTION OF THE DRAWINGS

The file of this patent contains at least one drawing executed in color. Copies of this patent with color drawing(s) will be provided by the Patent and Trademark Office upon request and payment of the necessary fee.

The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate an implementation of the invention and, together with the description, serve to explain the advantages and principles of the invention. In the drawings:

Figure 1 depicts an overview block diagram for use with methods and systems consistent with certain embodiments of the present invention when providing allele calls.

Figure 2 depicts a flow chart of the steps performed by a data processing system in processing allele calls when practicing methods and systems consistent with certain embodiments of the present invention.

SUB A1

Figures 3A-3D depict exemplary allele calling algorithms for use with methods and systems consistent with certain embodiments of the present invention.

Figure 4 depicts a flow chart of the steps performed by the committee machine of Figure 1 for use with methods and systems consistent with certain embodiments of the present invention.

Figure 5 depicts a block diagram of a system for practicing methods and systems consistent with certain embodiments of the present invention.

Figure 6 depicts data that may be generated and then interpreted using certain embodiments of the present invention.

Figure 7 depicts data discussed in Example II (Envelope Caller).

Figure 8 depicts data discussed in Example III (Optimizer Caller).

Figure 9 depicts methods for searching for an allele that is discussed in Example III (Optimizer Caller).

SUB A2

Figures 10 through 12 depict data that can be evaluated with the heuristic algorithm according to certain embodiments.

Figure 13 illustrates a typical standard heterozygous allele signature. (Circles denote user annotated allele calls. x-axis is in base pairs. y-axis is in A/D counts (voltage intensity))

Figure 14 illustrates steps in the allele calling routine according to the embodiments

discussed in Example V (Committee Machine Processing). First the signal is simplified via sampling and its peaks are located. This forms the target signal that is to be approximated. The two interconnected boxes indicate the process of varying the parameters and testing how closely the resulting signal matches the sampled version of the original. The set of parameters that yield the closest match contain the allele calls.

Figure 15 depicts data discussed in Example V (Committee Machine Processing). It illustrates hypothesis formation in the optimizer routine. The two columns represent the optimal solution (left column) and a suboptimal solution (right column). Panel (a) shows the target vector with the two red lines showing the location of the candidate peaks. Panel (c) shows the hypothesis formed using different values of stutter and $^{+}$A. Panel (c) shows the residual error resulting from subtraction of the signal in panel (c) from the signal in panel (a) (sum squared error = 0.0355). Panels (b,d,f) show the same process for a slightly different allele hypothesis. This is a poor hypothesis and the residual is rather significant (SSE = 0.4715). The x-axis is somewhat meaningless at this point since it gets mapped back to base-pair indices after the winning hypothesis is chosen.

Figure 16 depicts data discussed in Example V (Committee Machine Processing), and shows division of heterozygous signal into panels by the Envelope Caller algorithm. The panels are ranked according to signal energy and the three of interest are labeled p1, p2 and p3 with the two panels containing strong allele signatures being shaded in blue. Circles denote user annotated allele calls. (x-axis is in base pairs. y-axis is in A/D counts (voltage intensity))

Figure 17 illustrates and example of how reporting could be accomplished as discussed in

Example V (Committee Machine Processing). These are examples where consensus was not reached and show data that is difficult to interpret.

DETAILED DESCRIPTION

The following detailed description of the invention refers to the accompanying drawings. Although the description includes exemplary implementations, other implementations are possible, and changes may be made to the implementations described without departing from the spirit and scope of the invention. The following detailed description does not limit the invention. Instead, the scope of the invention is defined by the appended claims. Wherever possible, the same reference numbers will be used throughout the drawings and the following description to refer to the same or like parts. Several documents are discussed throughout this application. All of those documents are expressly incorporated by reference herein in their entirety for any purpose.

The following definitions are provided for terms used in this application.

Allele - An allele is one of two or more alternate forms of the same locus. With respect to a given locus, a diploid organism may be homozygous (having the same allele on each of the two homologous chromosomes) or heterozygous (having a different allele on each of the two homologous chromosomes). Non-diploid organisms may have more than two alleles.

Allele Calling – When fragment analysis is performed, the region of nucleic acid containing the marker is flanked by known primer sites which permit localization of the allele. For example, changes in the allele may result in different fragment lengths. Thus, for these

alleles, determination of the length of the nucleic acid sequence between primers is referred to as allele calling. For example, if two alleles are present, there will be two pieces of nucleic acid with different lengths.

Locus – A unique chromosomal location defining the position of an individual nucleic acid sequence.

Allele Signature – During PCR amplification, PCR stutter often occurs, which results in additional peaks that emerge in a predictable pattern. Another artifact that may appear is plus A distortion. The combination of the original signal, the stutter, and other artifacts is referred to as the allele signature.

Marker – Markers can be thought of as landmarks in the genome and can appear in noncoding regions of nucleic acid. Their use in linkage mapping stems from their polymorphism. Many different types of markers exist.

The following description involves allele calling when one analyzes dinucleotide repeats at given loci using PCR amplification. The invention is in no way limited to such work and may involve any number of repeats or may involve other types of genetic polymorphisms. Other polymorphisms include, but are not limited to, SNP's (single nucleotide polymorphisms), single base insertions and deletions, insertions and deletions involving more than one base, and rearrangements.

Similarly, embodiments of the algorithms may be applied to other types of data in which multiple algorithms produce results that typically require interpretation and scoring in terms of their confidence values. Such other areas of application include, but are not limited to, the

following: basecalling (de novo, mixed base and comparative sequence); SNP basecalling; spot-finding for microarrays; protein sequencing; protein/gene expression; peptide searches (a noisy time series alignment problem); and modeling of biological systems. One skilled in the art will appreciate all of the many types of nucleic acid and amino acid information that may be evaluated according to the present inventions. Examples include, but are not limited to, data from any of the applications above and any evaluation of properties including nucleic acid or amino acid length, molecular weight, or nucleic acid or amino acid identity.

In the committee approach for all of these applications of interpreting data, one uses the output of more than one algorithm rather than relying upon but one algorithm. Often, different algorithms may have various advantages over others depending on various conditions. The committee approach uses different algorithms to generate a meaningful confidence value on the correct interpretation of multiple data points. According to certain embodiments, the committee approach is particularly powerful when combined with the concept of establishing the operating environment first, an example of which is illustrated by the Envelope Caller described herein.

To determine given alleles at various loci, one can use PCR to selectively amplify regions of the gene that are known to have different alleles. In this example, one attempts to locate different length dinucleotide repeats at given loci. U.S. Patent No. 5,580,728 describes certain methods that can be used according to the present invention to amplify the genetic material in a sample and to obtain data that correlates to the different lengths of amplified nucleic acids. U.S. Patent No. 5,580,728 and all documents cited therein are expressly incorporated by reference herein. Possible data that may be generated is shown in Figure 6.

Figure 6 illustrates results that include artifacts created by the PCR amplification process. Without such artifacts, that data would show peaks at 93 and 103 basepairs, which would indicate that the individual is heterozygous for the two alleles of size 93 and 103 basepairs. PCR stutter, however, introduces additional peaks at 91 and 89 for the allele at 93, and at 101, 99, and 97 for the allele at 103. The stutter results in fragments that are shorter by one or more dinucleotides than the actual allele in the sample. Also, during the PCR process, additional A nucleotides may be added, which results in artifacts in Figure 6 having an extra basepair (i.e., at 94 for the allele at 93 and at 104 for the allele at 103). Figure 6 shows a relatively simple pattern that represents a heterozygous individual with alleles 93 and 103 and that includes artifacts. The artifacts that may be introduced, however, are not always simply disregarded when the actual alleles are closer together and allele signatures overlap. Thus, the present invention provides systems for interpreting data and making correct allele calls.

PCR stutter and the addition of A nucleotides is discussed in U.S. Patent No. 5,580,728. That patent discusses a particular algorithm that may be used to try to make correct allele calls. The present invention provides typically more reliable allele calling. The present invention includes not only new algorithms, but also systems that use more than one algorithm to increase the reliability of the call.

Figure 1 depicts an overview block diagram of a committee system 100 in which methods and systems consistent with the present invention may be implemented. Data 102 includes typical size-called data from a DNA sequencer such as the ABI 3700 DNA sequencer (Applied Biosystems). Data 102 may be passed to multiple allele calling algorithms, such as the Envelope

Detection Caller algorithm 104, Optimizer Caller algorithm 106, and a Heuristic Caller algorithm 108. Envelope Detection Caller algorithm 104 detects a heterozygous allele pattern when alleles are well separated spatially. Optimizer Caller algorithm 106 identifies impulse functions (e.g., location of the allele peaks) given a response signal (e.g., a raw microsatellite signal). Heuristic Caller algorithm 108 uses multiple rules and filters to eliminate peaks that are not alleles from consideration. More information on algorithms 104, 106, and 108 is provided below.

Each algorithm reports their results to a committee machine 110 that uses logic and/or rules to assign a confidence level to the call. Committee machine 110 produces robust results and can accurately predict calls. That is, machine 110 receives call results from several callers and can provide a degree of confidence for the resulting calls based on a statistical probability of an answer being correct given the degree of consensus between the different callers. More information on the committee of experts is further described below. The confidence level may be created by considering agreement between calling algorithms 104, 106, and 108. Results 112 contain the confidence level for each test performed by committee machine 110, and results 112 are transmitted to a user of a computer 114.

The committee system 100 provides a number of benefits over traditional allele calling algorithms. First, since each algorithm uses a different strategy in determining whether there is a call, if all the callers agree, then an extremely high value of confidence may be placed on the calls. If, however, not all allele calling algorithms agree, differing levels of confidence may be placed on the calls depending upon which algorithms agree. By considering the level of

agreement between the different algorithms over a large population of data, statistically significant confidence values can be assigned to the allele calls.

I. System Operation

Figure 2 depicts a flow chart of the steps performed by a data processing system in processing allele calls according to certain embodiments. First, the data processing system receives size-called fragment analysis data (step 202).

The received data may then be processed using various allele calling algorithms (step 204). Each caller algorithm operates well in different environments and on different signals. By using more than one caller on the same set of data, committee machine 110 may assign a confidence level to the call. Algorithms may either examine the data's complexity, and should it pass certain requirements, make the appropriate calls or make the calls regardless of data complexity. Several exemplary calling algorithms are described in Figures 3A-3D.

Once the data is analyzed with each allele calling algorithm, the results of each call are transferred to a committee machine 110 (step 206). Committee machine 110 processes the results of the calls (step 208) and arbitrates a decision and assigns an appropriate confidence value for the results of the calling algorithms. The results of this arbitration are reported to a user as the fragment lengths (calls) accompanied by a confidence value (step 210).

II. Envelope Caller

Figure 3A depicts a flow chart of the steps performed by a data processing system when processing alleles with the Envelope Caller algorithm according to certain embodiments. The Envelope Caller algorithm typically is used to detect a heterozygous allele pattern where the alleles are well separated spatially. The Envelope Caller assesses the complexity of a signal from the nucleic acid sequencer prior to making a call and if the signal's complexity is below a threshold (i.e., the signal is in the caller's operating region) it makes the call. Thus, since the caller operates in a constrained region where it knows it stands an excellent chance of being correct, the call is extremely accurate.

First, the algorithm may perform preprocessing such as smoothing (step 302). For example, the algorithm may use N-point smoothing replacing each point with a local average over itself and N points on either side. By replacing each point with such a mean, noise is removed from the signal, and a smoother signal remains.

Next the minima and maxima of the signal are determined (step 303) using a technique such as the Savitzky-Golay algorithm (See, e.g., Numerical Recipes in C: The Art of Scientific Computing, William H. Press, Saul A. Teukolsky, WIlliam T. Vetterling, Brian P. Flannery, Cambridge University Press, 1992, pages 650-655) which uses calculation of the derivatives of the signal in its processing. Other peak detection methods may be used. This step reduces the signal's dimensionality significantly by effectively expressing the signal's general shape with fewer points. The effect of this can be seen in Figure 7. Here the original signal is the solid line. After calculation of the minima and maxima, the signal is represented as the broken line.

In step 304, a new signal is formed by retaining only the maxima. This has the effect of determining the envelope of the signal. In Figure 7, this signal is shown as the dotted line. Next, the signal is passed back through the algorithm that determines the minima and maxima (step 305). With this new representation the original signal is then divided into panels at each minimum (step 306). A panel is a large section of the signal that is bounded by the signal's deep local minima. In Figure 7, 6 panels exist and are bounded as outlined in table 1.

TABLE 1

Panel	Boundaries (basepairs)
1	80-97
2	97-110
3	110-112
4	112-115
5	115-123
6	123-130

In order to determine the signal complexity and whether or not the algorithm should make a call, the algorithm first determines if three panels exist (step 308). If, at least three panels exist, the algorithm computes an energy level for each panel, for example, by summing the square of each element in the panel (step 312). Other methods of assessing the signal's energy in defined regions may be used. Since the algorithm is searching for the envelope characteristic of two well separated alleles, one typically uses three panels to ascertain if two distinct allele signatures exist. When one is searching for X number of alleles, one typically uses X + 1 panels to ascertain if X distinct allele signatures exist.

Using the three largest energy levels (E1, E2, and E3, respectively – which in the figure correspond to panels 1, 2, and 5), the Envelope Caller algorithm performs a "threshold" determination (step 314). That is, using the three energy levels (E1, E2, and E3), the algorithm determines, for example in certain embodiments, whether E2 is greater than 20% of E1, and whether E3 is no more than 7% of E2. If these conditions exist in these embodiments, the signal is of sufficiently low complexity that the envelope caller can operate. The calls are then made by reporting the largest peaks in each of the panels with the greatest energy. Thus for the case illustrated in figure 7, the calls would be made at the peaks topped by the diamond symbol at 93 and 103 basepairs.

In summary, certain embodiments of the Envelope Caller may include the following:

1. Pass the signal through a min/max detection algorithm and discard the minima. Thus, an envelope of the signal is obtained by connecting the points that are maximal.

2. Pass this new signal through a min/max detection algorithm again.

3. Divide the signal into panels of interest using the min/max information. A panel of interest here is defined as one where the signal is initially low, then increases rapidly, and falls off again towards the baseline. In these embodiments, the energy in these regions is calculated by summing the squares of the data in these regions.

4. Consider only the three regions with the greatest energy.

5. Choose the two dominant peaks in the signal and that the signal represents a heterozygous condition. In such a case, the allele calls are the maxima in the two panels with the greatest energy.

The following code may be used according to certain embodiments of the Envelope Caller methods.

Line 6 calls the subroutine envelope (lines 21-53) which divides the signal into the panels and calculates the energy of the panels and then identifies the three panels with the greatest energy content. Line 10 tests the condition given in step 4. If these conditions are met, line 11 retrieves the allele calls.

```
1       d = fieldnames(D);
2       ind = [];
3
4       for i=1:size(d,1),
5               eval(['cur=D.' char(d(i)) ';']);
6                [A, h, p1, p2]=envelope(cur);
7
8       if size(A,1) > 3,
9               B(i,1:2)=[A(2,5)/A(1,5) A(3,5)/A(2,5)];
10               if (A(2,5)/A(1,5) > 0.2) & (A(3,5)/A(2,5) < 0.07),
11              [peak_ind height]=getEnvelopeCalls(A,cur.analyzed);
12              R(i).alleleList=[cur.analyzed(peak_ind,1) height'];
13              else
14              R(i).alleleList=[];
15              end
16      end
17 end
18
19
20      %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
21      function [A, h1, p1, p2] = envelope(cur, plotflag)
```

```
% function h1=divider(cur, plotflag)
%
% cur - structure containing the data
% plotflag - set to anything to plot the process
%
% h1 - vector of division points
%
%
anal = cur.analyzed;
f = peak_trough(anal(:,2)); % first pass through the min/max detector
p1 = [f.maxvals]; p2=[f.maxinds];
g = peak_trough(p1); % second pass through min/max detector
h1 = anal(round(p2(round(g.mininds))),1);
ind = [1 closest(cur.analyzed(:,1), h1) length(anal)];
for i=1:length(ind)-1,
A(i,1:2) = ind(i:i+1);
A(i,3) = diff( ind(i:i+1) );
A(i,4) = diff( [anal(A(i,1),1) anal(A(i,2),1) ]);
A(i,5) = sum(anal(A(i,1):A(i,2), 2).^2);
A(i,6) = A(i,5)/A(i,4);
end
[p ind]=sort(A(:,5));
A=A(flipud(ind),:);
if exist('plotflag'), plotDivisionLines(cur,A,h1,p1,p2); end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [peak_ind, height] = getEnvelopeCalls(A, cura)
for i=1:2,
        [height(i) peak_ind(i)]=max(cura(A(i,1):A(i,2),2));
        peak_ind(i)=peak_ind(i)+A(i,1);
end
```

III. Optimizer Caller

U.S. Patent No. 5,580,728, which is incorporated by reference, describes allele calling via deconvolution. This is similar to the Optimizer Caller algorithm consistent with certain embodiments of the present invention.

According to certain embodiments the Optimizer Caller operates as follows. The algorithm operates on the principle of deconvolution that identifies the impulse functions (location of the allele peaks) given the response signal (the raw microsatellite signal). The routine uses model fit optimization to effect deconvolution. The model parameters optimized are peak location, peak height, and stutter percentage.

According to certain embodiments, the algorithm first performs dimensionality reduction by sampling at bins and then identifies the largest peak as the dominant allele. Bins are locations where one would expect to find alleles. Due to the way the data is generated, fragment lengths seldom are reported as integer base pairs. Thus, any peak falling within some threshold of the center of the bin is said to be that length. In certain embodiments, this threshold is +/- 0.15 basepairs. Thus, if a peak were to be size-called at 100.87 basepairs and a bin existed at 101 bp, the peak would be reported as 101 bp.

Sampling at the bins allows one to eliminate data points from the analysis. Bins are determined by previously compiled data. For example, one may pass to the system an original set of bins based on previously compiled statistics that reflect expected allele locations, and a sampling grid is formed by interpolating a one basepair grid that accomodates these bins. This creates a continuum of bins spaced at one basepair intervals upon which the signal is sampled.

Through building models where the amount of stutter is varied, the algorithm selects the next most likely allele by choosing the impulse function whose model results in the lowest residual error when subtracted from the original signal.

The flowchart in Figure 3(B) according to certain embodiments illustrates the concept as follows:

1) Sample at the bins (320) – as discussed above, the bins are locations where one would expect to find alleles. Thus, the signal above is sampled at these locations. Typically these locations includes minima and maxima but will also contain other portions of the signal (flat regions, stutter peaks).

2) Find minimas and maximas (322) - using the Savitsky-Golay approach, the precise location of the minima and maxima are located. The maxima represent possible alleles.

3) Select dominant peak as one allele (324) - typically, the largest peak is an allele - selecting this peak is a safe strategy, the problem is now reduced to finding the other allele (it if is present).

4) Form a series of hypotheses (models) by varying the location of the secondary peak and the amount of stutter in both the dominant and secondary peaks (326).

5) Subtract each model from the signal found in step (2) (328). The residual is kept in a table.

6) Select model with the lowest residual (330) - the model that results in the lowest residual best describes the signal from step (2) and thus is declared the winner. The allele calls are the location of the alleles that resulted in the model.

7) Transmit calls to user after application of any additional rules (332) such as removing left peaks below a certain threshold - experimentation has shown that peaks below a certain threshold are usually noise.

According to certain embodiments, the main Optimizer Caller algorithm steps are summarized as follows:

1) Data Reduction:

Using the a priori bins passed in, a sampling grid which includes additional bins is constructed. Then the signal is sampled to give a simplified discrete representation of the microsatellite signal, essentially the peak heights at the centers of the bins. See Figure 8.

2) Find the highest peak and assume it is one of the allele peaks, the "A" allele. See Figure 8.

3) Search for the B allele:

The algorithm searches for the location, height, and stutter percentage of the B allele peak that minimizes the residual signal, that is, the signal left over after subtracting the hypothesized signal from the observed signal. (The B peak may in fact be the same as the A peak, i.e. a homozygote.)

Figure 9 illustrates two different attempts in the search for the B allele. Recall that the A allele has been assumed to be the highest peak. Different hypotheses for the location, height, and stutter percent for the B allele peak are made. A composite signal is generated by superimposing the A and B hypotheses. The hypothesized signal is then compared to the observed signal and a residual error is calculated. The hypothesis with the lowest residual error is reported as the B

allele.

The method used to search for the best B allele parameters is flexible. In the first implementation of this algorithm, simple heuristics were used to prune the search space, but it was essentially an exhaustive search for the best B allele. Methods such as conjugate gradient, simplex or simulated annealing could be applied.

IV. Heuristic Caller

Figure 3C depicts a flow chart of the steps performed by a data processing system when processing alleles with the Heuristic Caller algorithm according to certain embodiments. The Heuristic Caller algorithm uses multiple rules (filters) to eliminate peaks that are not alleles. By removing the peaks using the filters, the remaining peak(s) may be alleles.

First, any of a number of preprocessing steps may be performed. Examples include the N-point smoothing mentioned in the Envelope Caller or noise quantification (or Noise Checker). Noise quantification is used to assess the quality of the signal. An example of Noise Quantification includes:

1) taking the signal;

2) performing smoothing as in 302 of Figure 3A;

3) subtracting the smoothed signal from the original signal; and

4) summing the squares of the difference between the two signals to get the sum squared error (SSE).

If the signal is relatively noise free, the SSE will be low and more faith can be placed in

the calls. If the SSE is high then the user is alerted that it might be wise to look at the signal and make calls manually.

After any such preprocessing steps according to certain embodiments, the process includes step 342 where the Heuristic Caller algorithm forms a peak list using a peak detection algorithm such as the Savitzky-Golay algorithm. According to certain embodiments, a list is formed with an entry for each peak that contains the following three pieces of information; peak location, peak height, and peak width. Next, various filters are applied to remove peaks that are not the correct allele calls (step 344).

Nonlimiting examples of one or more rules that may be employed include:

Remove split peaks (Split peak checker)

Remove background peaks (Background peak checker)

Remove peaks due to plus A distortion (Plus A Checker)

Remove spikey peaks (Spike peak checker)

Remove shoulder peaks (Shoulder peak checker)

Remove stutter peaks (Stutter checker)

Split peaks are two peaks that appear in the peak list that are similar in height (for example, at least about 70%) and typically less than about 0.1 basepairs apart. They typically are caused by a mixture of double and single stranded DNA. According to certain embodiments, if split peaks are detected, only the highest of the split peaks is preserved.

Background peaks are spurious peaks that do not have any significant stutter. Stutter almost always occurs for dinucleotide markers. Thus, peaks that do not have any significant

stutter are considered background peaks and are removed from the list. Background peaks are typically due to sample contamination.

Spikey peaks are spurious peaks that are tall but have a width that is not typical of the other peaks. A peak list has height, width and location data. Thus, an average peak width can be determined and any peaks that are too narrow compared to the rest of the population are removed. They are typically caused by sample contamination.

Shoulder peaks are peaks that appear very close to another peak and thus have the appearance of a shoulder. They are similar to spikey peaks except typically are lower in height, greater than 0.1 bp away, and less than 1 bp away. These are often caused by instrument noise. In certain embodiments, the shoulder peaks are removed.

According to certain embodiments, the filters applied in step 344 include at least one of those shown in the flow chart of Figure 3D. The One basepair Checker checks neighboring peaks to see whether there are one basepair peaks present. In certain embodiments, one may change the order of the filters. For example, according to certain embodiments, the Plus A checker and the Shoulder peak checker are switched with one another in the flowchart of Figure 3D. (The Final Assembler shown in Figure 3D assembles the final results and calls the alleles.)

Once all non-allele peaks are removed the Heuristic Caller algorithm determines if there are one or two remaining peaks (step 346). If there are more than two remaining peaks, additional filters are applied (step 348) in order to reduce the number of peaks to one or two. These rules are based on special cases that have been determined via observation. An example of a rule would be when four peaks remain – generally, the lowest two can be removed. Once only

one or two peaks remain, they are designated as the allele calls and are passed to the committee machine (step 350).

Figures 10 through 12 depict data that can be evaluated with the heuristic algorithm according to certain embodiments.

V. Committee Machine Processing

The following examples A and B illustrate the Committee approach according to certain embodiments of the invention.

Example A

Figure 4 depicts the steps performed by committee machine 110 according to certain embodiments when determining the final allele calls to be reported to the user and their associated confidence values. Committee machine 110 arbitrates the calls by using a set of rules. An exemplary rule table (Table 2) is depicted below. First committee machine 110 determines which callers are in agreement (step 402).

Next, committee machine 110 determines the correct calls to transmit and assigns a confidence level for these calls (step 404). According to certain embodiments, the confidence level is determined by considering the various cases in Table 2 over a large sample set that is representative of typical data. For example, if all three algorithms are in agreement (case 1), the committee machine assumes that the call is 99.9% correct and thus assigns a confidence value of 0.999. If there is no call for Envelope caller, and the same call for the Optimizer and Heuristic callers, committee machine 110 defines the confidence value as 0.970. If there is no call for the Heuristic algorithm, and the same call for the Envelope method and the Optimizer, committee machine 110 passes those calls to the user and assigns a confidence value of 0.621. If only the Optimizer produces a call, committee machine 110 assigns a confidence value of 0.692 correct. And finally, any cases that do not fit into the above scenarios are assigned the calls given by the Heuristic algorithm and are assigned a confidence value of 0.771. The above listed determination of agreement is exemplary. One skilled in the art will appreciate that other determinations of confidences are available. For example, additional algorithms may be used to produce more accurate confidence levels according to certain embodiments.

TABLE 2

Results from callers	Confidence
Same call by all three algorithms	0.999
Same call by Optimizer and Heuristic Algorithms No call made by Envelope Caller	0.970
Same call by Envelope Caller and Optimizer No call made by Heuristic	0.621
Only the Optimizer calls	0.692
Any cases that do not fit into above categories are called by the Heuristic Algorithm	0.771

Confidence levels can also be assigned by a person who is familiar with use of the particular algorithms used in a committee approach and the results obtained. Drawing on their experience with the particular algorithms, such a person can assign confidence levels for each of the possible combined results that can be obtained by the various algorithms.

<u>Example B</u>

1. Allele Calling Algorithms

In this embodiment, three different allele calling algorithms are implemented. Each possesses a distinctly differently philosophy. The callers are

envelope: Only classifies heterozygous data below a level of complexity. It does so with an extremely high level of accuracy and uses a visual approach based on detection of the characteristic envelop of a relatively noise-free, strong heterozygous signal with good separation between the alleles. If the data looks problematic, envelope refuses to make a call.

optimizer: Uses a maximum likelihood approach involving the formulation of hypotheses based on parameterization of an allele signal using allele location, amount of stutter and +A

artifact. The hypothesis that best explains the signal's energy content is declared the winner and the allele calls are those used in forming the winning hypothesis.

heuristic: A rule-based system of allele calling. Initially, all peaks are designated alleles and expert rules are used to remove false candidates until only the true alleles remain.

A section devoted to each method follows.

a. Heuristic Caller

Certain programs implement Genotyper allele calling algorithm (ABI PRISM™ Genotyper® 2.0 User's Manual. PE Applied Biosystems, 1996. 850 Lincoln Centre Drive, Foster City, CA 94404) and reuse this algorithm for trinucleotide and tetrinucleotide markers during allele calling processes. The steps involved in the process are outlined below.

1. Locate peaks. Find and identify all peaks in the marker size range.

2. Label peaks. Declare all peaks alleles.

3. Global cutoff. Find the maximum peak. Any peak lower than a threshold is removed from the called alleles list. This threshold is determined as *cutoffValue* * the peak's maximum height where *cutoffvalue* is a user defined parameter.

4. $^{+}$A removal. For any two neighboring peaks, if the distance between the peaks is within a certain number (user parameter +A *distance*) and the ratio between the upstream peak's height and the downstream peak's height exceeds the user parameter $^{+}$A ratio, the downstream peak is deleted from the called alleles.

5. Stutter removal. For any neighboring two peaks, if the distance between the peaks is

within the user parameter stutter distance and the ratio between the downstream peak's height and the upstream peak's height is exceeds the user parameter stutter ratio, the upstream peak is deleted from the called alleles list.

6. Declare alleles. Any remaining peaks are declared to be alleles.

Figure 13 illustrates a typical standard heterozygous allele signature. (Circles denote user annotated allele calls. x-axis is in base pairs. y-axis is in A/D counts (voltage intensity))

The algorithms behave relatively well for clean dinucleotide marker data and very well for tetrinucleotide marker data. For trinucleotide markers, however, there is a lack of data and it is not known for sure how this algorithm will behave. In all likelihood however, it will probably perform very well.

Certain embodiments of this algorithm include five parameters: *cutoffValue*, $^{+}$A *distance*, $^{+}$A *ratio*, *stutter distance* and *stutter ratio*. The program provides default values for these parameters and allows the user to adjust these values in the User Interface.

In reviewing large amounts of dinucleotide marker data, it became evident that several situations existed where the Genotyper algorithm was not optimal. These situations constituted the vast majority Genotyper errors. These cases are

1. Differential amplification. One allele is much higher than another allele. The global cutoff rule removes the lower allele.

2. 1bp allele. Two alleles exist being separated by only one base pair.

3. Bleedthrough (pullup) peak. Peaks exists due to strong neighboring color peaks and multicomponenting inaccuracy. This may be less than optimal for HID applications.

4. Background peak. One single background peak exists due to poor gel slabs.

5. Spiky stutter peak. Abnormally high and narrow stutter peaks.

The heuristic algorithm addresses these potential sources of error.

The heuristic algorithm includes additional rules. According to certain embodiments, these rules use the combination of feature variables (peak height, peak width, peak begin position, peak end position, peak begin height, peak end height, peak height ratios among peaks, base pair intervals among peaks) to figure out which peaks should be called alleles. In certain embodiments, the algorithm proceeds as follows.

1. Noise Checker. The noise level in the signal is checked. If the signal is too noisy, the process is interrupted.

2. Split Peak Checker. The neighboring peaks are checked for splitting. If splitting exists, only the higher peak is preserved.

3. Background Peak Checker. The peaks are checked to see whether they are single background peaks.

4. Small/Shoulder Peak Checker. Insignificant peaks and/or shoulder peaks are removed.

5. Spike Peak Checker. Spikey stutter peaks are removed

6. $^{+}$A Checker. The $^{+}$A peaks are removed.

7. Stutter Checker. The stutter peaks are removed.

8. Special Peak Checker. The peaks are checked to see whether there is differential amplification.

9. Preferential amplification, or if one basepair alleles exist.

These additional rules perform very well and reduce the number of errors substantially.

b. Optimizer Caller

This calling strategy in this embodiment rests on the assumption that a reasonable model for an allele's signature can be used to build an approximation to the original signal. This approximation is then subtracted from the original signal. The estimate that yields the lowest residual error gives the location of the allele(s).

In examining allele signatures, PCR stutter and $^{+}$A distortion modify what would ideally be isolated peaks. These, coupled with noise, make locating alleles peak problematic. Figure 13 illustrates their effect on the signal. Here, PCR stutter appears as a series of diminishing peaks to the left of the main signals at 212 bp and 223 bp and the $^{+}$A distortion appears as a small peak on the right of the main lobes.

Assuming that the PCR stutter peaks decrease at a constant percentage and assigning a value to the $^{+}$A distortion, a simple model of the allele signature is parameterized using the following three pieces of information:

- allele location;
- allele height;
- percentage stutter.

Thus, a search space is created where one considers all combinations of these parameters for a series of candidate allele peaks and obtains their resulting images. These images may then be subtracted from the original signal and the set of parameters with the lowest residual is

considered the winner. In this way, the allele locations are identified. The process according to these embodiments is flowcharted in Figure 14.

In these embodiments, preprocessing simply involves sampling the original signal to reduce its dimensionality. This can be performed by calculating the most important features of the signal; the peaks and valleys. By representing the signal in such a compact form, the search space is reduced significantly. The peaks form the set of candidate allele peaks that will be considered as possibilities for the allele calls. After the preprocessing, the next two boxes show the varying the parameters and the calculation of the residual. This process is iterated, and in the final box, a winning set of allele peaks (it could be a set of one peak) is declared. Actual output of the algorithm is contained in Figure 15.

The frames presented here demonstrate two cases; the first (frames (a, c, e)) being the optimal solution and the second (column formed by frames (b, d, f)), shows a solution that while close, does not explain the signal very well and leaves a high residual error. In both cases, the top frame show the signal that is being approximated. The candidate alleles are given by the position of the red lines. The middle frames show the hypothesized signal given different stutter parameters. And finally, the bottom frames show the resulting residual. The column of images on the right clearly demonstrates a better hypothesis and thus is declared the winning hypothesis. Allele calls are given by the locations of proposed peaks (red lines).

c. Envelope Caller

The Envelope caller is developed on the principle that while the other callers will

generally make a call no matter what, the envelope caller will only call alleles if it determines that there is a high probably that it will be correct. It is extremely accurate when it makes a call. This boosts the confidence in the calls and removes an entire class of data from requiring further consideration. Its basis is in considering the envelope of the signal and should two large masses of energy be detected (two large humps in the signal), the data is determined to be heterozygous. Allele calling is then simply performed by finding the maximum peak in each hump. While some simple heuristic rules could be added to slightly increase the accuracy. Specifically, these could cover the handful of cases where mistakes are made. However, in certain embodiments, these additional heuristics typically are omitted and instead, the combination of all callers is used to increase confidence to the close to one hundred percent mark in this subset of the data. In certain embodiments, the calling strategies should be fundamentally different in order that they each display strengths for particular data and thus the addition of heuristic rules to this caller may cause it to lose its identity in such embodiments.

The process is illustrated according to certain embodiments in Figure 16. The signal has been broken into 6 panels and the energy calculated. Panels marked $p1$ and $p2$ are shaded to indicate that they contain the most energy. Energy is denoted E and is the sum of the signal squared. The panel marked $p3$ contains the third largest energy content. In certain embodiments, the algorithm proceeds to make a call if the following two criteria are met

$$(1) \qquad \frac{E_{p2}}{E_{p1}} > 0.2$$

$$(2) \qquad \frac{E_{p3}}{E_{p2}} < 0.07$$

The call is made by finding the maximums in each of panels 1 and 2. The values of 0.2 and 0.07 in equations 1 and 2 were determined via trial and error and appear to give a good separation between easily classified data and more ambiguous cases.

2. Combination strategy

In certain instances, the individual algorithms may not be optimal when employed alone. In the committee of experts approach, the degree of confidence for a call is based on the statistical probability of an answer being correct given the degree on consensus between the different callers. This is a particularly apt approach when one considers that one of the callers according to this embodiment only makes a call if it considers it justified. In this embodiment, data falls into one of the following five categories.

Same call for envelope, optimizer, heuristic: The three algorithms are in agreement. This leads to a highly reliable result.

Envelope fails to call, optimizer and the heuristic agree: The signal has been deemed to be more difficult to classify and the process is left to the two more sophisticated approaches. The result is shown to be quite reliable however it is somewhat less confident than above particularly for "bad" data.

Heuristic failed to call, others agree: Sometimes, the heuristic algorithm will not call. This is particularly true in the case of noisy data. In such cases, when agreement between

Envelope and the optimizer occurs, that result is presented and the confidence value is defined as the probability that such situations are correct.

Only the optimizer calls: This covers the situation where the data is so problematic that neither Envelope nor the heuristic algorithm calls.

Any data not previously called: Should data not be called in the above cases, it is passed to the heuristic routine for calling. Experiment has shown that this algorithm typically surpasses the optimizer in terms of its accuracy when working in isolation.

Results

Results on two series of data from different labs is given in Table 3.

TABLE 3

	Lab 1			Lab 2			Lab 3		
strategy	examples	correct	conf	examples	correct	conf	examples	correct	conf
same R1, R2, R3	44.2	99.99	0.999	24.6	99.9	0.999	26.1	99.99	0.999
no R1, same R2 R3	51.3	99.40	0.994	58.8	97.2	0.972	70.0	99.69	0.997
no R3, same R1 R2	0.00	0.00	na	0.5	62.1	0.621	0.2	89.29	0.893
only R2 calls	0.04	66.7	0.667	0.8	69.1	0.691	0.3	80.00	0.800
straglers R2	4.5	21.2	0.212	15.2	30.9	0.309	3.5	39.67	0.400
straglers R3	4.5	73.6	0.736	15.2	77.1	0.771	3.5	38.45	0.385

Table 3: Results illustrating the ladder of confidence values that is created by considering agreement between the calling algorithms. R1 - envelope, R2 - optimizer, R3 - heuristic. All columns are percent-ages except for *conf. examples* - percentage of examples in full data set that belong to the category *strategy*, the column correct gives the percentage of examples in that category that are

correct. conf is the confidence value, it is percentage correct for a given category. Total number of traces examined: Lab 1 - 10724, Lab2 - 8000, Lab 3 - 14192.

All numbers (except the confidence values) are percentages. The column labeled *examples* is the percentage of the data set that has fallen into that category. The next two columns recount the percentage of the data from column one that has been correctly and incorrectly classified. The percentage correct has been passed to the column *conf* to be used as a confidence value. One other casual observation is that lab two possesses data that is distinctly more difficult to process. This can be seen by the number of examples that have fallen through to the final level of processing. This data is marked *straglers*.

The final two rows are for the same chunk of data. They show that the default caller should be the heuristic as it has a higher percentage of correct calls.

Another interesting opportunity is to pass these results on to the customer as a report - particularly in the case of examples that have fallen into the "difficult to classify" category where no consensus exists. This could be in the form of Figure 17 and would provide a good visual aid for data checking. Figure 17 illustrates 25 markers, and though in some cases it appears that consensus was reached, it is not marked as such because the threshold to determine the "sameness" of calls was set too low. In most of the cases however, it can be seen why the data is problematic. The red circles give the user annotations while the three levels of asterisks give the calls for envelope, the optimizer, and the heuristic from bottom to top.

Conclusion

The multi-caller approach is significant in that it provides hard numbers for the confidence in the calls. As well, by partitioning the data into different categories based on how easily the data is classified, it does well in providing a method for checking results.

It is very important to keep in mind that the three methods should not be considered as competing. Rather, as they are based on entirely different philosophies, they serve to confirm each other. The heuristic caller has a vast amount of domain knowledge behind it. The optimizer employs a more formal detection and estimation framework whereby the hypotheses are formed about the allele locations and similar to maximum likelihood, the hypothesis that best explains the signal's energy is chosen as the most likely explanation. Envelope employs a very simple visual inspection to identify easily classified data. These three algorithms each have their strengths and when working in concert form a very robust system and the high degree of trust it is able to place in a call is by virtue of the fact that high confidence requires consensus from a variety of perspectives.

VI. Architecture

Figure 5 is a block diagram that illustrates a computer system 500, according to certain embodiments, upon which embodiments of the invention may be implemented. Computer system 500 includes a bus 502 or other communication mechanism for communicating information, and a processor 504 coupled with bus 502 for processing information. Computer system 500 also includes a memory 506, which can be a random access memory (RAM) or other

dynamic storage device, coupled to bus 502 for determining allele calls, and instructions to be executed by processor 504. Memory 506 also may be used for storing temporary variables or other intermediate information during execution of instructions to be executed by processor 504. Computer system 500 further includes a read only memory (ROM) 508 or other static storage device coupled to bus 502 for storing static information and instructions for processor 504. A storage device 510, such as a magnetic disk or optical disk, is provided and coupled to bus 502 for storing information and instructions.

Computer system 500 may be coupled via bus 502 to a display 512, such as a cathode ray tube (CRT) or liquid crystal display (LCD), for displaying information to a computer user. An input device 514, including alphanumeric and other keys, is coupled to bus 502 for communicating information and command selections to processor 504. Another type of user input device is cursor control 516, such as a mouse, a trackball or cursor direction keys for communicating direction information and command selections to processor 504 and for controlling cursor movement on display 512. This input device typically has two degrees of freedom in two axes, a first axis (e.g., x) and a second axis (e.g., y), that allows the device to specify positions in a plane.

A computer system 500 provides allele calls and provides a level of confidence for the various calls. Consistent with certain implementations of the invention, a level of confidence for an allele call is provided by computer system 500 in response to processor 504 executing one or more sequences of one or more instructions contained in memory 506. Such instructions may be read into memory 506 from another computer-readable medium, such as storage device 510.

Execution of the sequences of instructions contained in memory 506 causes processor 504 to perform the process states described herein. Alternatively hard-wired circuitry may be used in place of or in combination with software instructions to implement the invention. Thus implementations of the invention are not limited to any specific combination of hardware circuitry and software.

The term "computer-readable medium" as used herein refers to any media that participates in providing instructions to processor 504 for execution. Such a medium may take many forms, including but not limited to, non-volatile media, volatile media, and transmission media. Non-volatile media includes, for example, optical or magnetic disks, such as storage device 510. Volatile media includes dynamic memory, such as memory 506. Transmission media includes coaxial cables, copper wire, and fiber optics, including the wires that comprise bus 502. Transmission media can also take the form of acoustic or light waves, such as those generated during radio-wave and infra-red data communications.

Common forms of computer-readable media include, for example, a floppy disk, a flexible disk, hard disk, magnetic tape, or any other magnetic medium, a CD-ROM, any other optical medium, punch cards, papertape, any other physical medium with patterns of holes, a RAM, PROM, and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave as described hereinafter, or any other medium from which a computer can read.

Various forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to processor 504 for execution. For example, the instructions may initially be carried on magnetic disk of a remote computer. The remote

computer can load the instructions into its dynamic memory and send the instructions over a telephone line using a modem. A modem local to computer system 500 can receive the data on the telephone line and use an infra-red transmitter to convert the data to an infra-red signal. An infra-red detector coupled to bus 502 can receive the data carried in the infra-red signal and place the data on bus 502. Bus 502 carries the data to memory 506, from which processor 504 retrieves and executes the instructions. The instructions received by memory 506 may optionally be stored on storage device 510 either before or after execution by processor 504.

As explained, systems consistent with certain embodiments of the present invention provide a committee machine that receives calls as input from at least two different allele calling algorithms. By receiving these calls, the committee machine is able to determine a level of confidence in a variety of conditions.

The foregoing description of an implementation of the invention has been presented for purposes of illustration and description. It is not exhaustive and does not limit the invention to the precise form disclosed. Modifications and variations are possible in light of the above teachings or may be acquired from practicing of the invention. For example, although the foregoing was primarily described with reference to particular allele calling algorithms, the concepts of the invention could also be applied to any other type of allele calling algorithms, such as *TrueAllele* from *Cybergenetics* or the *Genetic Profiler* program from *Molecular Dynamics*. When different algorithms are used, one can assign confidence values for the possible combination of results as discussed above, e.g., by analyzing various cases over a large

sample set that is representative of the data or by having a skilled person familiar with the algorithms assigning such confidence values based on experience. Additionally, the described implementation includes software but the present invention may be implemented as a combination of hardware and software or in hardware alone. The invention may be implemented with both object-oriented and non-object-oriented programming systems.